A Study on Test-Time Augmentation and Attention Mechanism in DeepLabv3+ for Deep Learning-Based Segmentation

Sy-Phuc Pham[1], Hyung-Jeong Yang[1], Duy-Phuong Dao[1], Soo-Hyung Kim[1],
Guee-Sang Lee[1]
Corresponding Author: hjyang@jnu.ac.kr
[1]Dept. of Artificial Intelligence Convergence, Chonnam National University, South Korea
{phamsyphuc123,duyphuongcri}@gmail.com,{shkim,gslee}@jnu.ac.kr

CEUR-WS.org/Vol-3181/paper50.pdf

ABSTRACT

In this paper, we present an approach for segmenting polyps in a two-dimensional image. In this challenge, we conducted expermients with four different methods and developed a segmentation model based on DeepLabv3+ and attention mechanism. Besides, we applied the standard augmentations for the data preprocessing and the test-time augmentation technique for improving the prediction mask. Our ensemble models have been evaluated on the development dataset, which was provided by the MediaEval challenge.

1 INTRODUCTION

In the Medico task, it was a challenge to explore the approaches to automatically segmenting images collected from the human colon. In this task, we took part in subtask 1: Polyp segmentation. The goal of this subtask was to develop algorithms for segmenting polyps in endoscopy images. With two-dimensional (2D) medical images, we have to develop a deep learning model for segmenting the polyp. We propose three main models to do the Medico task. For the first model, we apply DeepLabv3+ [6].This is the encoder-decoder architecture popular in the segmentation area, with the strength of the Atrous Spatial Pyramid Pooling in the encoder path. In the second model, we use U-Net-MobileNetv2, which was popular in medical image segmentation. Based on the U-Net architecture, the encoder path has been developed using MobileNetv2. The last model is the proposed model, which we detail in section 3.

The content of this paper is organized as follows. Section 2 presents the summary of some existing research closely related to polyp semantic segmentation. In section 3, we discuss the proposed method. Details of the experiment and results are presented in section 4. Finally, the conclusion is presented in section 5.

2 RELATED WORK

The U-Net [15], a popular architecture in biomedical image segmentation consists of the encoder-decoder architecture with the advantages of skip connection. The authors in [16] used the U-Net model with the augmentation of Kvasir-SEG dataset [10]. The Kvasir-SEG dataset includes 1000 samples which were collected from endoscopy images. Marcus et al. in [2] developed the U-Net by using pre-trained MobileNetv2 in the encoder path, and used Adadelta optimizer to segment the polyp in the Kvasir-SEG. Based on the model in [2], Saruar et al. [1] changed MobileNetV2 to ResNet50 in the encoder path and kept the decoder path. The authors in [12] did the polyp segmentation in the MediaEval2020 challenge via Self-Knowledge Distillation [9] with the ResUNet++ [11] backbone. Multi-SuperVision Net [14], an encoder-decoder architecture with five layers was used in the MediaEval2020 challenge. In [14], the encoder path has been kept similar to the encoder path in U-Net, and the decoder path has been developed using a combination of dense blocks and Concurrent Spatial and Channel Attention.

3 PROPOSED METHOD

In this section, we discuss the architecture of the proposed method DeepLabv3+ with Self-Attention (DLV3SA). It contains two components: DeepLabv3+ [6] and Self-Attention mechanism [17], as shown in Figure 1.

DeepLab based on the encoder-decoder architecture, has four versions: DeepLabv1 [3], DeepLabv2 [4], DeepLabv3 [5], and DeepLabv3+ [6]. But in our approach, we used the DeepLabv3+ architecture as it is the most recent version with high performance in segmentation tasks. The feature extraction from ResNet [8] was transmitted into a multi-layer Deep Convolutional Neural Network (DCNN). The DCNN creates a feature map, a spatial representation of the features of the input image. In the DCNN, Atrous convolutions [6] are used for feature filtering instead of Convolutional Neural Network (CNN) as the field of view of filters in Atrous convolution is wider than CNN, and it doesn't reduce the dimension of the feature map too deeply. However, it still retains the number of parameters and computational cost.

The Attention mechanism is the important layer in the encoder path of the proposed model. The network takes advantage of the attention mechanism's capabilities to concentrate the features with more meaningful data and to reduce the number of feature maps. We employed an Attention layer on the output of each layer of the Atrous Spatial Pyramid Pooling (ASPP) to focus on important information. By assigning a weight to each layer in ASPP, the attention layer can get the best output before concatenating the four outputs of ASPP. We can represent the attention layer as shown in Equation 1.

$$Attention_{f_i}(Q^{f_i}, K^{f_i}, V^{f_i}) = softmax(Q^{f_i} {K^{f_i}}^T) V^{f_i} \qquad (1)$$

where f_i represent the feature extraction from ResNet, Q is the query, K is the key, V is the value, and $Q = K = V$ when this is the self-attention.

Copyright 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
MediaEval'21, December 13-15 2021, Online

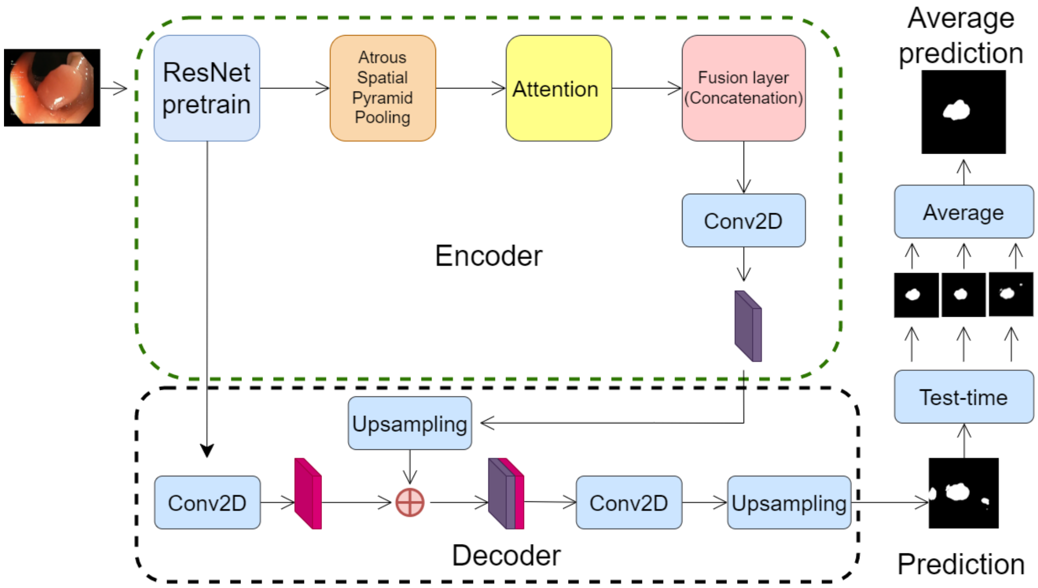

Figure 1: Proposed model.

Table 1: Our results on private test

STT	Model	Accuracy	Jaccard	Dice	F1	Recall	Precision
Submission 1	DeepLabv3+	0.9482	0.6394	0.7383	0.7383	0.7122	**0.853**
Submission 2	DLV3SA with augmentation	**0.951**	0.6577	0.7469	0.7469	**0.747**	0.8102
Submission 3	U-Net-MobileNetV2	0.9445	0.6364	0.7332	0.7322	0.7416	0.7977
Submission 4	DLV3SA with test time augmentation	0.9489	**0.659**	**0.7563**	**0.7563**	0.7402	0.8469

Test-time augmentation (TTA) [13] is a technique for improving the prediction mask. It involves creating multiple augmented copies of each image in the test set, having the model make a prediction for each augmented copies, and then returning an ensemble of those predictions. To apply TTA, we used fliplr and flipud functions from Numpy library [7]. We calculated the average of all of the masks and got the final result for submission.

4 EXPERIMENT AND RESULTS

In the subtask 1 of the Medico task, for submission 1, we used the DeepLabv3+. We ran the model on five different train-validation splits with the ratio 8:2 (80% for training and 20% for validation). For submission 2, we applied data augmentation such as HorizontalFlip and VerticalFlip in data preprocessing and split the dataset with a ratio of 8:2 and trained with the DLV3SA model. For submission 3, we trained the U-Net-MobileNetV2 model with a dataset with more epochs than the number of epochs in submission 1 and 2. In the last submission, we ran the experiment with the proposed method discussed in section 3.

In subtask 1, we receive the results from the organizer in six metrics: accuracy, Jaccard, dice, F1, recall, and precision, as shown in Table 1. In submission 1, the Jaccard score is better than in submission 3. With the DLV3SA model, the Jaccard score in submission 4 is better than in submission 2. Table 1 shows that our proposed model outperforms DeepLabv3+ and U-Net-MobileNetV2 in terms of accuracy, Jaccard index, Dice similarity coefficient, and F1 of 0.9489, 0.659, 0.7563, and 0.7563, respectively, on the official test dataset.

5 CONCLUSION

In this paper, we present the proposed model for automatic polyp segmentation. In our work, the data augmentation technique is applied to input images in submission 2. We also applied the Attention mechanism in the encoder path of DeepLabv3+ to improve the outputs from the ASPP module. Our experimental results show that our proposed model outperforms U-Net-MobileNetV2 and DeepLabv3+ on the unseen test set. In the future, we plan to experiment with more than one multiple pre-trained models in the encoder path by fusing their feature maps before putting them into the ASPP module.

ACKNOWLEDGMENTS

This research was supported by the Bio Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961).

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1A2B5B01002085).

REFERENCES

[1] Saruar Alam, Nikhil Kumar Tomar, Aarati Thakur, Debesh Jha, and Ashish Rauniyar. 2020. Automatic Polyp Segmentation using U-Net-ResNet50. *arXiv preprint arXiv:2012.15247* (2020).

[2] Marcus VL Branch and Adriele S Carvalho. 2021. Polyp Segmentation in Colonoscopy Images using U-Net-MobileNetV2. *arXiv preprint arXiv:2103.15715* (2021).

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* (2014).

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV).* 801–818.

[7] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, 7825 (Sept. 2020), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[10] Debesh Jha, Pia H Smedsrud, Dag Johansen, Thomas de Lange, Håvard D Johansen, Pål Halvorsen, and Michael A Riegler. 2020. A Comprehensive Study on Colorectal Polyp Segmentation with ResUNet++, Conditional Random Field and Test-Time Augmentation. (2020).

[11] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. 2019. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM).* IEEE, 225–2255.

[12] Jaeyong Kang and Jeonghwan Gwak. 2020. KD-ResUNet++: Automatic Polyp Segmentation via Self-Knowledge Distillation. In *MediaEval 2020 Workshop.*

[13] Nikita Moshkov, Botond Mathe, Attila Kertesz-Farkas, Reka Hollandi, and Peter Horvath. 2020. Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific reports* 10, 1 (2020), 1–7.

[14] Sabari Nathan and Suganya Ramamoorthy. 2020. Efficient Supervision Net: Polyp Segmentation Using EfficientNet and Attention Unit. In *MediaEval 2020 Workshop.*

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention.* Springer, 234–241.

[16] Shruti Shrestha, Bishesh Khanal, and Sharib Ali. 2020. Ensemble U-Net model for efficient polyp segmentation. In *MediaEval 2020 Workshop.*

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems.* 5998–6008.